DOCUMENT RESUME

ED 413 347 TM 027 655

AUTHOR Mayer, Daniel P.
TITLE Will New Teaching Standards Be Implemented If Old Tests Are the Yardstick for Success?
SPONS AGENCY College Entrance Examination Board, New York, NY.
PUB DATE 1997-03-00
NOTE 47p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Algebra; Educational Assessment; Educational Practices; Grade 8; *Junior High School Students; Junior High Schools; *Mathematics; Performance Factors; *Standardized Tests; *Standards; Teaching Methods; *Test Results
IDENTIFIERS Longitudinal Study of American Youth; National Assessment of Educational Progress; National Council of Teachers of Mathematics; *NCTM Professional Teaching Standards

ABSTRACT

As almost every state revises its mathematics frameworks or develops new ones, current testing practices remain almost unchanged. The National Council of Teachers of Mathematics (NCTM), in its "Professional Standards for Teaching Mathematics," argues for greater emphasis on application, reasoning, and conceptual understanding, but the "Standards" authors recognized that a discontinuity exists between standardized tests and their view of mathematics teaching. The implications of this mismatch may be that the "Standards" are never really implemented. This study examines whether students taught in NCTM-like classrooms perform differently on standardized assessments than students taught in traditional classrooms. The question is addressed in the context of eighth-grade algebra classrooms using students from the second cohort of the Longitudinal Study of American Youth for the 1988-89 school year. The analytic sample consists of 325 students and 37 teachers from 34 schools. A teacher survey determined whether teachers used an active teaching approach of the sort advocated by the "Standards." Algebra achievement was measured through test items from the National Assessment of Educational Progress (NAEP) taken in eighth and ninth grades. The study finds that the more emphasis a teacher places on class discussion and small group work relative to lecture and seatwork, the less students gain on their NAEP examinations over 1 year. This negative association is not accounted for by teacher or student background characteristics. Results highlight the mismatch between the sort of teaching increasingly advocated and performance on current types of standardized tests. (Contains 2 figures, 4 tables, and 61 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 347

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Daniel Mayer
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Will new teaching standards be implemented if old tests are the yardstick for success?

Daniel P. Mayer
Mathematica Policy Research, Inc.

Paper presented at the 1997 Annual Meeting of the American Educational Research Association in Chicago.

BEST COPY AVAILABLE

TM027655

Research for this paper was supported by The College Board. Opinions and conclusions are those of the author and do not necessarily reflect the views of the supporting agency. Please send comments to Daniel Mayer at 7215 Holly Ave., Takoma Park, MD 20912. E-mail: dmayer@mathematica-mpr.org.

The latest wave of ambitious education reforms in the United States may be undermined by a discontinuity: As almost every state either substantially revises their mathematics frameworks or develops new ones (Blank and Pechman 1995), current testing practices remain largely unchanged (Blank et al. 1995). The frameworks reflect a "profound and unprecedented shift" away from policies which tinker at the edges of the educational process toward those which focus on its heart. "In its two and quarter centuries, the United States has never had explicit education content or performance goals" (Marshall, Fuhrman, and O'Day 1994, p. 12), but the standards movement is changing this. Even the extensive reform efforts of the 1970s and 1980s remained aloof from curriculum and teaching practices. During those decades, policy makers tried to improve schooling by adjusting resource allocations (e.g., striving for racial balance and financial equity) and by setting outcome goals (e.g., minimum course requirements and minimum competency tests). The perceived failures of these policies has led to our country's current obsession with educational standards. State curriculum frameworks, the federal government's Goals 2000 initiative, and the rapid-fire succession of new standards documents produced by the professional curriculum associations are all testaments to the prominence of this movement.

Most of the mathematics curriculum frameworks contain explicit recommendations regarding teaching practices that are heavily influenced by the National Council of Teachers of Mathematics' *Professional Standards for Teaching Mathematics* (Blank and Pechman 1995). This should come as no surprise as this council was one of the earliest and most important players in the development of curriculum and teaching standards. The ideas presented in these *Standards* undergird most curriculum frameworks and other prominent science, mathematics, and technology education reform movements in the United States and other developed countries (Black and Atkin 1996).

The NCTM *Professional Standards* argue that optimal learning of mathematics requires that teachers place less emphasis on memorization of facts and mastery of routine skills and greater weight on application, reasoning, and conceptual understanding. The *Standards* state that for students to "understand what they learn, they must enact for themselves verbs that permeate the mathematics curriculum: 'examine,' 'represent,' 'transform,' 'solve,' 'apply,' 'prove,' 'communicate.' ***This happens most readily when students work in groups, engage in discussion, make presentations, and in other ways take charge of their own learning"*** (National Council of Teachers of Mathematics 1989, pp. 58-59, emphasis added).

The *Standard's* authors were clearly worried that a discontinuity existed between traditional standardized tests and their *Standards*. They warned in their earliest *Standards* document that "[i]n an instructional environment that demands a deeper understanding of mathematics, test instruments that call for only the identification of single correct responses ***no longer suffice***" (National Council of Teachers of Mathematics 1989, p.192, emphasis added). Since issuing this warning, a few states, and several school districts, have experimented with alternative assessments such as portfolios and open-ended question tests. But despite these signs of change, traditional assessments remain the norm throughout the country (Blank et al. 1995).

The implications of this mismatch may be that the *Standards* will never be implemented. "As achievement test results have become more and more influential in policy decisions, the degree of overlap between the content tested and the content taught has also increased in importance" (Wiley and Yoon 1995, p. 355). Wiley and Yoon noted Resnick and Resnick's guiding principles for accountability assessments, the first of which is "[y]ou get what you assess. Educators will teach to tests if the tests matter in their own or their students' lives." The second principle is that, "[y]ou do not get what you do not assess" (Resnick and Resnick 1991, p. 59).

A recent article in the *New York Times Magazine* illustrates why educational decision makers in the trenches may choose to ignore the new state mathematics frameworks. The article, titled "Scores Count," focuses on one New York City principal's efforts to help poor black students pass the state's exams. In commenting on why he pushes a back-to-basics curriculum rather than an NCTM-type approach, he notes the importance of the current testing environment: "...we can't just say that we don't believe in measuring our kids against kids in other parts of the city or state, because colleges are making decisions based on those comparisons. Employers are making decisions based on those comparisons...I have to make sure that my students are competitive. I'm dealing with what *is*. Harvard is still looking at scores, so that's what I'm going with" (Mosle 1996).

Are the authors of the *Standards* and this principal correct to assume that a mismatch exists? The results from extant research offer contradictory evidence. But, as shown below, the ambiguity exists because the research base is both thin and flawed. Given the importance of this issue, this paper hopes to provide more concrete information about the potential for a mismatch by answering the following question: *Do students taught in NCTM-like classrooms perform* ***differently*** *on standardized assessments than students taught in traditional classrooms?*

Because of a heightened interest nationwide in expanding access to algebra, this question is addressed within the context of 8th grade algebra classrooms. In the past few years, New York City, Philadelphia, Boston, Atlanta, Louisville, Milwaukee, and Oakland, to name but a few school districts, have succeeded in dramatically increasing their 8th grade algebra enrollments (Clark 1994; Hart 1994; White 1993). The impetus for expanding access to algebra comes from research which highlights the increased odds that students taking algebra early will take more advanced

mathematics in high school, have higher mathematics performance by the end of high school (Stevenson, Schiller, and Schneider 1994), and make it to college (Pelavin and Kane 1990).

Limitations of the existing research base

As noted above, the research base is both thin and flawed. It is thin, in large part, because most researchers have been interested in exploring whether the *Standards* ***work.*** Answering the effectiveness question is entirely different from exploring whether a mismatch between the *Standards* and traditional tests exits. Since the NCTM and many researchers argue that standardized tests are not a valid measure of the progress made in active teaching classrooms, answering the effectiveness question requires that students be measured with non-standardized tests. Thus, researchers interested in this question have either just described the process of reform (e.g. Ball and Schroeder 1992; Ferrini-Mundy and Johnson 1994) or they have developed special assessments for their studies (e.g. Campbell 1995; Hiebert and Wearne 1993) . Thus, these studies shed no light on the mismatch hypothesis.

A more limited number of studies exist which look at the relationship between NCTM-type teaching practices and standardized test scores, and these, obviously, are the ones which are important to this study's research question. The remainder of this section will therefore concentrate on the lessons learned from them.

While some argue that the teaching approaches which mirror those endorsed by NCTM lower standardized test scores (e.g. Hirsch 1996) others have found that they boost them (e.g. Knapp and Associates 1995). Hirsch bases his claims on findings from early "process-product research" (i.e. research which links classroom processes like teaching practice to products like test

scores), one of the most influential research traditions in the study of teaching (Brophy and Good 1986; Shulman 1986). While numerous studies point to the same conclusion, generalizations from these early process-product studies should be made with caution because the research was conducted prior to the development of the learning theories embedded in the *NCTM Professional Standards*.

The researchers who developed the NCTM learning theories provided useful insights about knowledge acquisition, but those researchers offer no empirical evidence concerning the relationship between the NCTM-endorsed teaching approaches and standardized, or even non-standardized, tests (e.g. Case and Bereiter 1984; Cobb and Steffe 1983; Hiebert 1986; Lampert 1986; Lesh and Landau 1983; Schoenfeld 1987).

A handful of recent process-product studies do directly test these learning theories using traditional standardized tests. Carpenter et al (1989), Cobb at al. (1991), Knapp and Associates (1995), and Simon and Schifter (1993) found that students were ***not*** penalized if their teachers used an NCTM-type approach in their classrooms. In fact, Knapp and Associates and Carpenter et al. found that the NCTM taught students ***outperform*** their counterparts in more traditional classrooms.

How relevant are these studies to this study, and how much faith should we place on the generalizability of their conclusions? On the one hand, their relevance is limited since none of them explicitly focused on secondary mathematics, let alone algebra, and two of them focused explicitly on the early elementary grades (e.g.Carpenter et al. 1989; Cobb et al. 1991). On the other hand, each of the studies did measure the effects of the NCTM teaching approach on standardized test scores. Even if we thought these studies were relevant, can we have faith in their conclusions? All four have design limitations which will be explored below.

For example, neither Simon and Schifter (1993), nor Cobb et al. (1991), established the validity of their study's most important variable, teaching practice. Their findings rely on the assumption that real differences exist between the teaching strategies employed in the treatment and control groups. Simon and Schifter (1993) establish what type of teaching occurred in the classrooms by using measures of student attitudes and beliefs toward mathematics. Common sense would suggest that this approach only provides a very rough approximation of teacher practice. The authors feed this skepticism by offering no information pertaining to the reliability or validity of their measures. Cobb et al. (1991) also use an indirect way of ascertaining what the teachers do in their classrooms. Though using teacher reports may seem to be like an effective method at first blush, the survey questions they use only ask about pedagogic ***beliefs, not practice***. Recent research suggests that when teachers discuss their teaching in the abstract it often fails to accurately capture what they do in practice (Burstein et al. 1995). Unfortunately, Cobb et al. offer no corroborating evidence to prove that beliefs and practice are correlated. If we cannot have faith in the measures of teaching practice, then we cannot be confident that these studies were correct in assuming that teaching style really differed between the control and treatment teachers.

These studies are also limited by their lack of attention to confounding factors which might drive their findings. Cobb et al. created control and treatment groups within each of three separate schools, but the teachers who taught in the treatment classrooms were a self-selected group. Each of the teachers in the treatment group ***volunteered*** to participate in a summer workshop and each volunteered to receive "extensive support" from the trainers throughout the school year. The control teachers and classes consisted of the other teachers (those who did not volunteer) in those same three schools. Were these teachers less interested in learning new teaching approaches? Were

they more senior? Less senior? Better educated? Less well educated? Of a certain gender or race? Raudenbush, Rowan, and Cheong (1991) argue that a teacher's background and training affects the probability that she emphasizes an NCTM-type approach in her classroom. Teachers with more years of teaching experience, higher levels of education, and a deeper knowledge of the content being taught in their classrooms may well be more likely to emphasize an active teaching approach. On the other hand, one could argue that new teachers are more open to experimenting with new teaching approaches and they will be more likely to use the NCTM approach. Either way, regardless of their teaching style, it is likely that more experienced teachers would be more effective instructors. Therefore, ignoring the profiles of the teachers is a major oversight if one wants to truly isolate the impact of teaching style.

Simon and Shifter used a design which controls for differences in teachers, yet their study has other limitations. These researchers used a historical design where the teachers received the intervention in the middle of the study. The first year of the study teachers taught in their typical fashion. Then, over the summer, before the second year of the study, the teachers volunteered to receive training in the new teaching approach. After the second academic year the test results from the students from year one were compared to the test results from the students from year two. Were there differences between the year one and two students? The literature on educational achievement clearly indicates that student characteristics play an important role in determining performance on achievement tests. Female students (e.g. Benbow and Stanley 1981; Fennema and Sherman 1977) and low socioeconomic status students (e.g. Coleman and et al. 1966; Jenks et al. 1972) may perform worse on mathematics achievement tests than their male and more well-off peers. On the other hand, given findings from evaluation of science curriculums, it is possible that low SES students will respond more positively to the NCTM intervention than higher SES students (Bredderman 1983). Simon and Schifter

did not investigate these issues and thus leave us wondering whether their findings may be driven by them.

Another potential confounding factor that both Cobb et al. and Simon and Shifter ignore is that teachers use different teaching approaches depending on the overall makeup of their classrooms. Metz (1978) and Oakes (1985) found that classrooms comprised of more advanced students are provided with more opportunity to engage in critical thinking, which may in turn mean that their teachers use a more active teaching approach. How were students sorted into the classrooms in both of these studies? Were the more advantaged students placed in classrooms with teachers who were more likely to use an NCTM approach? If this was the case, to what degree did this affect their conclusions?

Knapp et al. (1995) and Carpenter et al. (1989) successfully avoided some of the pitfalls which plague Cobb et al. and Simon and Shifter's research. They defend their teaching practice measures by offering validity and reliability data. Knapp et al. used a longitudinal design to look at the natural variation in teaching styles in high-poverty elementary school classrooms. They utilized teacher logs, classroom observations, and teacher surveys to establish the validity and reliability of their measure and to prove that variation in teaching practice existed. Carpenter et. al. (1989) randomly assigned first grade teachers into treatment and control groups and then validated that the workshops the treatment teachers attended created a difference between the two groups in terms of their pedagogical approaches.

These studies also deal directly with the confounding factors cited above. Knapp et al. statistically controlled for important student, teacher, and classroom level conditions. Carpenter et al. use the more powerful technique of randomly assigning treatment and control groups.

Both studies found evidence that students taught by an NCTM teaching approach performed better on standardized tests than students taught in a more traditional manner. Knapp's found that the students taught in NCTM-type classrooms outperformed the other students on both the arithmetic computation and the mathematical concepts and applications portion of the Comprehensive Test of Basic Skills. Carpenter et al. found that the treatment classes marginally outperformed the control classes on the word problems portion of the Iowa Test of Basic Skills, but they found no difference in performance between the treatment and control groups on the computational skills portion of the test.

Knapp and Carpenter's studies have their own limitations, however. These authors, as well as Cobb et al. and Simon and Shifter, use pre-test scores to help them measure student knowledge at one point in time, even though the limitations of this approach have been well documented (Willett, 1994). First, this approach does what it's supposed to do poorly. Most researchers use the pre-test to control for differences in the students' initial status. However, a pre-test can only imperfectly control for initial differences and this leads to biased parameter estimates (Rogosa, Brandt, and Zimowski 1982). A second source of bias comes from the correlation between the pre-score and any unobserved influences on student achievement, such as SES or parental involvement (Willett, 1994). Thus, by controlling for initial status with a pre-test, these authors use a measure of achievement which probably reflects more about student background characteristics than it does about student learning. To avoid this problem, researchers should study ***changes*** in achievement over time by at least creating a gain score (Rogosa and Saner 1995).

Finally, all four of these studies are flawed in their analyses.. Because students are nested within classrooms, it is likely that there are unobserved student characteristics within each

classroom which are highly correlated with one another. It has been well documented that this situation, left unattended, results in biased estimates of the parameters' standard errors (e.g., Bryk and Raudenbush 1992). Thus, these authors create some doubt about what they proclaim is, and is not, statistically significant.

The above discussion reveals that little direct evidence exists about the potential for a mismatch between the NCTM-recommended teaching approach and traditional assessments, and that the little that does exist is filled with design flaws (ignoring selection bias, using poor measures of teaching practice and student learning , and ignoring the implication of having students nested within classrooms). This study, and future research in the field, needs to build upon these previous studies by accounting for their omissions. These concerns are addressed in methodology section of this study, which is described next.

Methods

Sample

This analysis uses students from cohort 2 of the Longitudinal Study of American Youth during the 1988-1989 school year (Miller et al. 1991). This longitudinal study of students enrolled in a national probability sample of 52 junior high schools is well-suited for answering my research question because it explicitly focuses on the mathematics and science experiences of middle school students and includes a rich array of data. The only other recent national probability sample with information on the 8th grade classroom experience, NELS:88, lacks LSAY's focus on mathematics, and also lacks a critical variable, a pre-8th grade achievement measure.

The analytic sample comes from the 473 LSAY 8th grade Algebra 1 students. These students studied with one of 49 teachers at 43 separate schools. Missing data from variables used in

this analysis reduced the sample to 325 students, and 37 teachers from 34 schools.[1] A two sample t-test comparing the means of students with missing data to the students without missing data across non-missing measures established that the two groups were not statistically different from one another on any student level measure used in this analysis, including student gain scores ($p=.14$), socioeconomic status ($p=.37$), gender ($p=.30$), and parental involvement ($p=.49$). (For definitions of these measures see the measures section below.) The number of sampled students per classroom teacher in this sample ranges from 1 to 28, with the median of 7.

Given Metz (1978) and Oakes' (1985) findings that classrooms made-up of more advanced students are given more opportunity to engage in critical thinking, we need to examine whether selectivity bias plays a role in this data. Since the majority of students take algebra after 8th grade, this study's whole sample comes from the academically elite and this should minimize the academic differences between classrooms. Over 74 percent of LSAY 8th graders were excluded from this study because they were in "lower" mathematics classes. In addition, the mean fall mathematics test score for the students included in this study was significantly higher than those students' in lower classess (63.5 versus 50.8, $p<.001$). Finally, and perhaps most importantly, the teachers perceive these students as being among the elite. Thirty of the 37 teachers (three did not respond to this question) rated the algebra students at "the highest [mathematics] ability level," and 34 teachers ranked these students as "somewhat higher than average." Thus, it seems unlikely that the students prior ability will drive the results of this study. However, to be certain of this, we will explore in the results section below whether there is a relationship between teaching style and the student pre-test scores.

Measurement of Teaching Style

To assess the degree to which the algebra teachers utilized an active teaching approach of the sort endorsed by the *Standards*, a composite of four variables was created. The research cited above suggests that active teachers spend time acquiring knowledge about their students' cognitive processes and simultaneously engaging students in activities which promote an opportunity for them to construct their own meaning of the material being covered. Conversely, passive teachers treat students as "empty vessels" and impart information to them in a rote, teacher directed manner.

Teacher self-reports reveal how much classroom time LSAY algebra teachers dedicated to two passive activities, lecture and seatwork, and two active approaches, teacher-lead discussions and small group work. The validity of using self-reported survey data pertaining to instructional strategies is supported by Porter et al. (1993) and Burstein et al. (1995). Porter et al. reported "substantial" agreement between survey responses pertaining to instructional style and detailed teacher logs describing actual lessons. They concluded that their "validation results were very encouraging" (A-5). Burstein et al. used an even more rigorous statistical test to examine the agreement between survey and log data and found that the survey data "report accurately the instructional strategies used most often by teachers" (p.45).

The survey questions used in this study are very similar to those used by Porter et al. and Burstein et al. Teachers were asked to approximate the amount of time devoted to different instructional approaches. The response options were: none, 30 minutes, one hour, two hours, more than three hours a week.

Frequency distributions indicate that these teachers spent more time engaged in passive teaching activities (see Table 1). The mean responses reveal that the greatest proportion of class

time was spent on seatwork and lecturing. Fifteen teachers reported spending more than 2 hours a week lecturing, while only 3 teachers spent more than 2 hours a week involved in either small group work or discussion.

Table 1 Here

Correlations among these four variables ranged from .19 to .51 (with a median of about .35). These low to moderate correlations suggest that teachers do not exclusively utilize a passive or an active approach, but rather employ both approaches to varying degrees.[2] The Cronbach's Alpha coefficient revealed that a moderate to strong composite could be created (alpha=.65) and a principal component analysis confirmed this.[3] Since the weights from the first PCA composite were roughly equal (ranging from .45 to .57), the four variables were summed and used as an index of active teaching. This construct represents the degree to which a given teacher employs an active teaching style. The more active the teaching style the higher the score. Scores have been standardized to have a mean of 0 and a standard deviation of 1 and range from -2.5 to 2.1.

Measures of Algebra Achievement

The National Assessment of Educational Progress (NAEP) items employed by LSAY in their exams of student performance are used to measure the students' overall mathematics achievement. The items used by LSAY were developed by NAEP in 1986 to measure "development in skills in cognitive processes related to achievement in mathematics" (Miller et al. 1991, p. 56). The tests measured three distinct ability categories and seven subject matter domains. The ability categories are the following:

> (1) Skill and Knowledge in Mathematics refers to recall and recognition of mathematical words and symbols and their use in straightforward, routine manipulation leading directly to

> answers in a single or a very few steps. (2) Routine Application involves the use of mathematical knowledge and skill in solving problems that are similar to those the student would have encountered in textbook example and classroom assignments. (3) Problem Solving and Understanding requires interpretations of underlying concepts, assumptions, and relationships and their use in solving non-routine, often multi-step problems (Miller et al. 1991, p. 56).

The seven subject matter domains include: mathematical methods; discrete mathematics; data organization and interpretation; measurement; numbers and operations, relations; functions and algebraic expressions; and geometry.

The tests used a multiple choice format and were administered to the students in the fall of their 8th and 9th grade. Though the exams were designed to be completed in a 50 minute period, in most cases, students needing more time were given it.

Is this test a traditional multiple choice test? Though the NAEP items used for this study were constructed after much of the research cited in support of the *Standards* had been published, it does not appear that the 1986 NAEP utilized this research in their test construction (National Assessment of Educational Progress 1986).[4]

In fact, the 1986 NAEP embraced the "single correct response" approach which most states still rely on (Blank et al. 1995; Romberg, Wilson, and Khaketla 1989; Silver and Kenney 1993) and which the NCTM recommended moving away from. Consequently, this test suits our purpose of trying to assertain how students taught by teachers who use an NCTM type approach perform relative to students in other classrooms when compared using traditinal standardized tests?

Creating a Gain Score

Scores in the fall of 8th grade and the fall of 9th grade were estimated using Item Response Theory (IRT) methods and were scaled to allow for the measurement of growth over time. The

reliability of the 8th and 9th grade IRT-based scales were high (.91 and .92, respectively). They were estimated by taking the ratio of the variance of the observed scores to the estimated variance of the latent scores and were measured separately for the 8th and 9th grade exam.

A gain score was used rather than simply employing the pre-test as a predictor variable of the post-test. This option was more appealing than controlling for the pre-test for several reasons. First, the purpose of using the pre-test as a control variable would be to control for initial status. However, this approach only imperfectly controls for prior standing, thus resulting in biased parameter estimates (Rogosa, Brandt, and Zimowski 1982). A second source of bias embedded in using a pre-test control measure comes from the correlation between the pre-score and any unobserved influences on student achievement, such as SES or parental involvement. Third, since the students were not randomly assigned they may potentially come from two different populations and the distribution of their pretest scores could regress to different means. "Without a clear rationale for assuming the two populations are drawn from the same underlying population, it is safer to use a difference analysis rather than an ANCOVA [analysis of covariance] to measure change" (Laird 1992, p. 191). Finally, the interpretation of a gain score is much more meaningful since it measures the relationship between student ***learning*** and important predictors of change (in this case, most importantly, teaching style). Controlling for initial status with a pre-test only measures post-test ***achievement*** by imperfectly controlling for background characteristics in order to make all students appear to start out "equal" before taking the post-test.

The gain score used in this analysis ranges from -32 to 48. However, the middle 75 percent of the scores only stretch out from .85 to 8.5. There is a fairly symmetric distribution. There is a mean gain of 4.67 and a median of 4.28. The large standard deviation of 8.77 represents the fact that several students scored more than two standard deviations from the mean in

either direction. In fact, 4 values lie more than 2.5 standard deviations above the mean and 4 are 2.5 standard deviations below.[5]

Measures of Control Variables

Students

Student gender, socioeconomic status (SES), and the degree of family emphasis on academics were included in the analysis. Gender was self-reported by the students. SES and family emphasis are measures created by LSAY by compositing various student and parent survey responses. The SES composite combines parent-reported educational and occupational levels for the student's mother and father with student- reported information on the household's possessions. The parent academic push measure also combines student and parent reports. Students responded to eight checklist items which asked about their parents behavior. For example: My parents " tell me how proud they are when I make good grades," "insist I do my homework," and "reward me for getting good grades." The two parent questions used in this composite asked either the student's mother or father "how often do you talk to your son/daughter about how well he/she is doing in school?" and "how often do you talk to your son/daughter about homework and school projects?"

Table 2 provides descriptive statistics on these student measures. The SES composite ranges from a low SES of -1.3 to a high of 1.81, the average student has an SES score of .41 (SD=.68). The parental push composite ranges from a low of 1.3 to a high of 10.0 and is skewed toward lower values (mean=7.93, SD=1.79).

Table 3 illustrates that, at the student level, gender, SES, and parental academic expectations are weakly and insignificantly correlated with test score gains. This suggests that student-level characteristics may not explain any significant variation in the gain scores.

Not unexpectedly, these algebra students have higher gain scores and SES composites than their peers in the lower math classes. A two sample t-test reveals that the mean SES of -.03 and mean gain score of 3.23 for the non-algebra students is significantly different from the Algebra students' mean of .41 and 4.67 ($p < .01$, and $p < .01$, respectively). The parental expectations of these two groups are, however, similar (7.71 for the lower students versus 7.93 for the higher students, $p = .10$) .

Table 2 Here

Table 3 Here

Teachers

Teaching style, possession of a masters degree, the number of postsecondary mathematics courses taken, and the number of years teaching were investigated. Table 2 shows that 57 percent of the teachers had masters degrees, and on average they took almost 12 postsecondary mathematics courses (SD=5). The average teacher taught for 16 years (SD=7.31), but this ranged widely from 1 to 31 years. The correlations presented in Table 3 illustrate that surprisingly little association exists between teaching style and both the number of years of teaching experience and whether a teacher has a masters degree. However, the number of postsecondary mathematics courses is moderately negatively correlated with an active teaching style ($r = -.34$, $p < .05$). It appears that the more mathematics training a teacher has, the less likely she is to employ an "active" teaching

approach (i.e., the more likely she is to employ a traditional teaching style). If "active" teachers obtain better results in their classroom, this finding could have important implications regarding the preparation of mathematics teachers.

Just as the algebra students differed from the non-algebra students in important ways, so do their teachers. As Metz (1978) and Oakes (1985) suggested, the algebra teachers tended to employ a more active pedagogical style than the teachers who taught the "lower" mathematics classes (.12 versus -.03, $p < .02$). The algebra teachers had also taken more college mathematics courses (11.9 versus 9.1, $p < .01$) and had been teaching for more years (16.1 versus 14.7, $p < .01$).

Analytic Approach

Accurately measuring the impact of an active teaching style on test score gains not only requires that the confounding factors be controlled for, but also requires that two statistical problems presented by these data be attended to: the hierarchical nature of the data (i.e., students are nested within classrooms), and measurement error in the primary predictor. Left unattended, the first problem results in incorrect standard errors for the parameter estimates (Bryk and Raudenbush 1992), and the second leads to a parameter estimate for teaching style which is negatively biased (Fuller 1987). Because a parameter's significance is determined by dividing its estimated coefficient by its standard error, bias in ***either*** will produce an uncertain gauge of its significance. Consequently, to determine the significance and magnitude of the active teaching coefficient, both problems are confronted in the subsequent analysis. Though no one analytic approach is currently available to address both of these problems simultaneously, the use of two separate statistical approaches offers us valuable information about the significance and magnitude of the relationship.

Multilevel models

OLS statistical models do not account for the hierarchical nature of this data and yield biased estimates of the parameter's standard errors, while multilevel models can account for the clustering of students within classrooms and generate unbiased estimates (Bryk and Raudenbush 1992). Consequently, to explore the relationship between classroom processes and student mathematics test scores, multilevel models were used to examine the relationship between student gains scores on the one hand, and teacher and student characteristics on the other. These models linked student gains to classroom and student predictors using a pair of statistical models. The first of these models (the "level-1" model) expresses the gain in mathematics achievement for student i ***within*** classroom j during the academic year as a function of the student characteristics:

$$Gain_{ij} = \beta_{0j} + \beta_{1j}(SES_{ij} - \overline{SES}_{ij}) + \beta_{2j}(Male_{ij}) + \beta_{3j}(Push_{ij} - \overline{Push}_{ij}) + \varepsilon_{ij}$$

SES and parental push were centered around their respective class means in order to make the β_{0j} coefficient represent a female student's expected average gain within classroom j. The β_{1j}, β_{2j} and β_{3j} each represent, respectively, the relationship between SES and math gain in classroom j, gender and math gain in classroom j, and parent academic push and math gain in classroom j

The second of the two models (the "level-2" model) expressed the parameters from the level-1 model as a function of the teacher characteristics and investigated whether the parameters in the level-1 model differed ***across*** classrooms. It is that fact

--that the level-2 model expresses the within-class estimates from the level-1 model as a function of the classroom level predictors--which allowed me to addresses the primary research question. In initial analyses, the effect of all of the level-1 predictors (SES, gender, and academic push) were allowed to vary between classrooms. However, as will be explained in the next section, in the final

model these effects were fixed across classes. As a result, only the intercept, β_{0j} -- the average gain for students in a class -- varied across teachers, yielding a level-2 model that could be written as:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Style}_j - \overline{Style}) + \gamma_{02}(\text{Years}_j - \overline{Years}) + \gamma_{03}(\text{Masters}_j - \overline{Masters})$$

$$+ \gamma_{04}(\text{Math Training}_j - \overline{MathTraining}) + \nu_{0j}$$

The intercept from the level-1 model, β_{0j}, represents the main effects of teaching, while had the other coefficients continued to be treated as random, they would have represented interaction effects. Since all of the teacher characteristics (style, masters degree, experience) were centered around their grand mean, γ_{00} represents the average gain score for girls. γ_{01} represents the relationship between the average gain and the average teaching style and answers the primary research question by showing whether teachers with a higher "style" score (i.e., a more active teacher) have, on average, higher gain scores in their classes. γ_{02}, γ_{03}, and γ_{04}, represent the relationship between average gain and a teacher's years of experience, possession of a master's degree, and the number of postsecondary math courses taken.

The primary advantage of multilevel models is illustrated by the fact that these two equations allowed for the explicit inclusion of measurement error within and between classrooms (ε_{ij} and ν_{0j}). Since OLS regression does not account for the fact that the student level observations are clustered within the teacher level observations, the usual assumption of independent errors is potentially violated. Another advantage of hierarchical models is that it provided estimates of the total within and between teacher variance. This made it possible to estimate the amount of between-classroom variation that exists in gain scores and to measure how much of the variation could be "explained" by the classroom level variables used in this analysis.

Measurement Error Models

The problem with relying solely on multilevel models for this analysis is that, like their OLS counterparts, these models assume that all the predictors are measured without error. Because measurement error exists in the active teaching style variable, using multilevel models alone would have led to an underestimate of the magnitude of the variable of primary interest. Measurement error in an explanatory variable produces a negative bias in that variable's parameter estimate. In addition, depending upon the correlations among the predictors, this could either negatively or positively bias the estimated effects of other predictors in the model. An "errors-in-variables" correction can correct for measurement error by removing the measurement error associated with teaching style from the predictor-predictor covariance matrix (Fuller 1987).

To make this correction, single level regression models were fit which simultaneously included the classroom and student level variables and used the errors in variables correction procedure provided in the Stata software package (Stata Corporation 1993). The correction involved using the reliability estimate of active teaching obtained when the teaching style variables were composited (alpha=.65). Stata's "eivreg" procedure corrects the parameter estimates and goodness-of-fit statistics using the method described by (Fuller 1987).

Analytic Strategy

Because the reliability correction did not need to be used until the active teaching composite was included in the model, this analysis first explored the relationship between student and teacher background characteristics and student gain scores using multilevel models. Using HLM3 (Bryk, Raudenbush, and Congdon 1993) multilevel models were fit in four stages. First, in order to identify a baseline student-level model, only the student background predictors (gender, SES, and parent push) were included at level-1 and no predictors were included at level-2. If any of these

predictors were significantly related to test score gains, or if their relationship within classrooms varied significantly across classrooms, they were used in stage two of the analysis. In stage two, the teacher background predictors (years of teaching experience, possession of a master's degree, and the number of postsecondary mathematics courses taken) were added at level-2 to the baseline model to see if they explained any of the variation in the test score gains across classrooms and to see if they interacted with gender, the only significant level-1 predictor. In stage three, teaching style was added at level-2 to a parsimonious model including only the predictors with p-values of less than .10 from stage one and two. Finally, all variables excluded from the model during the first and second stages were reintroduced to the "final" model to see if their relationship with student test score gains changed because of the addition of the variables at later stages.

After arriving at a final HLM model, the same variables included in that model were fit into a final single-level reliability-corrected model and the results from the two statistical techniques were compared. Since both the HLM standard error and the reliability corrected parameter estimate associated with active teaching are unbiased; the following method was used to interpret these two separate "final" models: The magnitude of all of the HLM standard errors were compared to the non-reliability corrected OLS produced standard errors in order to provide an estimate of the bias associated with utilizing a non-reliability corrected linear model. This estimate was then used to inflate the standard error estimate associated with the reliability corrected parameter estimates so that an approximate unbiased estimate of each parameter's p-value could be obtained. Finally, since the magnitude of the reliability corrected parameter estimates is unbiased, these estimates were used to interpret the size of the effect of active teaching on gain scores.

Results

The degree to which test score gains vary by classrooms is of paramount importance because if no differences exists between the 37 classrooms, then there is nothing for teaching style to explain. However, as Figure 1 shows, the average gain score within each of the classrooms ranges from 11.92 to -9.87, with the classroom's average gain being 4.34. The Figure also shows that within classrooms the amount of variation is substantial, ranging from 20.21 to 0. An HLM analysis, with no covariates at either level-1 or level-2, reveals that a modest nine percent of explainable variation exists among these classrooms (variance=6.8; $\chi^2_{36}=68.10$; $p<.001$). This data clearly reveals that over the course of the year some students gained more on the NAEP exams if they were in particular classrooms. The question is, does the teachers pedagogic style explain any of this variation?

Figure 1 Here

Effects of student background. As the correlations foreshadowed, and Table 4 confirms, a model containing gender, SES, and parental involvement at level-1 and no variables at level-2 (model 1) reveals that none of these student background characteristics are significantly related to test score gains ***within*** classrooms. Significant variation between classrooms existed in the within classroom differential between males and females (variance=26.75; $\chi^2_{31}=57.00$, $p<.01$), but stage two of the analysis showed that this observed variation was not related to any of the teacher-level indicators used in this analysis. In addition, since none of the other student characteristics' relationship to gain scores became significant after adding level-2 covariates, they were excluded from the final model.

Effects of teacher background. Model 2 in Table 4 included the teacher background variables at level-2, and no level-1 variables. This model revealed that the academic preparation of teachers (possession of a master's degree and number of postsecondary mathematics courses) is not significantly related to gain scores, while the number of years teaching is positively related with a p-value less than .10.[6]

Table 4 Here

Effects of teaching style. In model 3, teaching style was combined with teaching experience at level-2, and no covariates were included at level-1. In this model, teaching experience is significantly positively related to student gain scores ($p < .05$) and teaching style is ***negatively*** related to gain scores, but with a p-value of less than .11. The inclusion of previously excluded predictors (i.e., student SES and gender, and the teacher training variables) did not alter this finding. In other words, student gains are higher in classrooms where teachers use less active teaching strategies. Model 3 indicates that students in classes in which teachers utilized an active teaching style had ***lower*** test score gains than their peers in classrooms in which teachers used a passive style. The more emphasis a teacher places on class discussion and small group work, relative to lecture and seatwork, the less students gained on their NAEP exams between the fall of their 8th grade and the fall of 9th grade.

A reliability corrected version of this model corrected for the bias in the parameter estimate of style. The HLM coefficient rose from -1.03 to -1.54 and its p-value dropped from less than .11 to .04 (model 4). In order to ascertain whether this reliability corrected estimate would be significant if we had an unbiased estimate of its standard error, the multilevel models correction

(explained above) was used. This correction increased the p-value associated with style from .04 to less than .10 (model 5).

The magnitude of this relationship is illustrated in Figure 2 which shows that controlling for teaching experience, a one standard deviation difference in active teaching style (sd=1) is predicted to be associated with a drop in student gain scores of 1.54. The standardized effect size of teaching style and teaching experience are both .18.[7] The figure illustrates that a novice teacher who utilizes a highly traditional pedagogic style (two standard deviations below the mean) could have students whose gain scores are, on average, greater than her more experienced colleagues who might utilize a more active pedagogic style.

Figure 2 Here

Can this unexpected negative relationship between teaching style and student test score gains be explained by modeling the data differently or by some kind of selection bias? It was hypothesized that since more experienced teachers produced better results in their classrooms, younger teachers might be so ***ineffective*** at using an active teaching style approach that the students in their classes had the negative "gain" scores. In other words, perhaps an interaction between teaching style and experience would explain away the findings. This was tested, but rejected ($p > .50$).

As far as selectivity bias, above it was hypothesized that the most able students would be more likely to be taught using active teaching methods. However, if it were actually the less able students who had the most active teachers, this might explain why those teachers had lower gain

scores on average. To check this hypothesis, a hierarchical linear model was fit regressing the fall 8th grade test score on a model which only contained the teaching style composite at level-2. Though the results were marginally insignificant ($p=.14$), the direction of the coefficient was supportive of the original hypotheses that the students in the active classroom were ***more*** ($\beta=1.21$)[8], not less, advanced than the students in the other classrooms. When these results are combined with the fact that no evidence exists to suggest a ceiling effect in the 9th grade test scores, the selectivity hypotheses does not seem to hold. In fact, these findings suggest that since the "advanced" students are in the most active classrooms and are gaining less on the NAEP exam than their less advanced peers, the magnitude of the relationship between teaching style and gain scores could in fact be underestimated.

Explanatory power of the model. Since the hierarchical linear model produces an estimate of the amount of residual (i.e., or potentially explainable) variance at both the student and teacher level, it is more informative to present these findings rather than the less precise R2 estimate from the reliability-corrected model. None of the within classroom variables (SES, gender, and parental expectations) turned out to be significantly related to gain scores, and therefore our final model did not explain any of the within-classroom variation. At level-2, however, a fairly large portion of the explainable variance was accounted for. The variance was 6.80 between classrooms before any variables were added and 3.97 in our final model. This shows that a teacher's pedagogic style and years of experience accounted for 42 percent of the residual variance between classroom gain scores.

Conclusion

This study found that the more emphasis a teacher places on class discussion and small group work relative to lecture and seatwork, the less students gained on their NAEP exams over one year. This suggests that students taught by teachers who emphasize an NCTM-type teaching approach may perform worse on traditional standardized tests relative to their peers in more conventionally taught classrooms. This ***negative*** association is not accounted for by teacher or student background characteristics: neither the academic preparation of the teacher nor the SES of the students are significantly related to the students' NAEP test score gains. The variation in gain scores between males and females within classrooms differed across classrooms, and years of teaching experience is positively associated with gain scores, but these factors do not interact with, or supplant, the negative effect of an active pedagogical style.

Furthermore, the magnitude of the relationship between teaching style and gain scores is not trivial. The association between a teacher's teaching style and her students' test score gains is as great as the association between a teacher's years of classroom experience and her students' test score gains.

These results create some complicated puzzles. Do students in the NCTM classrooms actually learn something that is not captured by traditional standardized test and makes up for the drop in test scores found in this study? If this were the case, it would lend support to NCTM's concern that current assessments do not "reflect the scope and intent of [their] instructional program" (National Council of Teachers of Mathematics 1989, p. 192).

Alternatively, is it possible that an active teaching environment, while necessary for students to ***begin*** to "construct their own mathematical meaning," ultimately may not be sufficient. The NCTM warned evaluators that:

> A high quality mathematics experience is not determined simply by the presence of computers or calculators or the use of small groups, manipulatives, or student discussions. The nature of the mathematical task posed and what is expected of students are critical aspects against which to judge the effectiveness of the lesson (National Council of Teachers of Mathematics 1989, p.5).

Although classrooms where teachers emphasize group work and discussion appear more active, teachers in those classrooms may not be truly pushing their students "[t]o understand what they learn [and]... enact for themselves verbs that permeate the mathematics curriculum: 'examine,' 'represent,' 'transform,' 'solve,' 'apply,' 'prove,' 'communicate'" (as quoted in the National Council of Teachers of Mathematics 1991, p.2). Unfortunately, the LSAY data do not reveal much about the instructional content used in the small groups, discussions, lectures, and seatwork.

Keeping in mind that the LSAY data predates the *Standards*, it is quite possible that teachers today understand better what is meant by active teaching and are therefore more effective at utilizing the instructional methods studied here. However, the negative findings in this study might foreshadow a recurring problem in education reform. Teachers often take "reform" initiatives into their classroom and reconstruct them into something that the reform creators might not recognize (Cohen 1990; Cuban 1984), which in turn may undermine the reform.

These findings are not anomalous. Although the early process-product studies predate the development of the NCTM learning theories, they suggested that students taught in a manner consistent with the NCTM approach would perform worse on standardized tests than more traditionally taught students. Later process-product studies found either no statistically significant difference between traditionally and non-traditionally taught students or a slight positive difference, but their research designs raise several questions about their generalizability.

Given the investment of energy required by teachers to master and implement a new teaching repertoire, policy makers must be concerned about the results of this study. Why would teachers or administrators push to implement the mathematics frameworks if they will have no effect on, or, worse, harm test scores?

If old assessments are the explanation for these negative results, then as long as they are used educators have little incentive to adopt the new teaching methods. Studies should be undertaken to investigate what kind of impact this mismatch has on teacher behavior. A comparative analysis looking at the level of adoption of the new teaching standards between states which use old assessments to those using new ones would tell us whether teachers are actually responding to the mismatch identified in this study.

Sources

Ball, D.L., and T.L. Schroeder. 1992. Improving teaching, not standardizing it: How do the professional standards connect to the curriculum and evaluation standards. *The Mathematics Teacher* 85 (1):67-68.

Benbow, C.P., and J.C. Stanley. 1981. Sex differences in mathematical ability: Fact or artifact? *Science* 210:1262-1264.

Black, P.J., and J.M. Atkin, eds. 1996. *Changing the subject: Innovations in science, mathematics, and technology education*. New York City: Routledge.

Blank, R.K., C. Hemphill, S.L. Sardina, D Langesen, and B Brathwaite. 1995. State education policies on k-12 curriculum, student assessment, and teacher certification: 1995. Washington, DC: Council of Chief State School Officers.

Blank, R.K., and E.M. Pechman. 1995. State curriculum frameworks in mathematics and science: How are they changing across the states? Washington, DC: Council of Chief State School Officers.

Bredderman, T. 1983. Effects of activity-based elementary science on student outcomes: A quantitative synthesis. *Review of Educational Research* 53:499-918.

Brophy, J., and T.L. Good. 1986. Teacher behavior and student achievment. In *Handbook of research on teaching*, edited by M. Wittrock. New York: Macmillan.

Bryk, A.S. , and S.W. Raudenbush. 1992. *Hierarchical linear models: applications and data analysis methods*. Newbury Park: Sage Publications.

HLM2 and HLM3 computer programs and users' guide .

Burstein, L. , L.M. McDonnell, J. Van Winkle, T. Ormseth, J. Mirocha, and G. Guitton. 1995. Validating national curriculum indicators. Santa Monica: RAND Corporation.

Campbell, F.P. 1995. Project IMPACT: Mathematics achievement in predominately minority elementary classrooms attempting reform. Paper read at Annual meeting of the American Educational Research Association, at San Francisco, CA.

Carpenter, T.P. , E. Fennema, P. Peterson, C.P. Chiang, and M. Loef. 1989. Using knowledge of childgen's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal* 26 (4):499-531.

Case, R., and C. Bereiter. 1984. From behaviorism to cognitive development. *Instructional Science* 13:141-58.

Clark, T. 1994. Interview, Director of Research for the Philadelphia Public Schools.

Cobb, P., T. Wood, E. Yackel, G. Wheatley, B. Trigatti, and M Perlwitz. 1991. Assessment of a probelem-centered second-grade mathematics project. *Journal for Research in Mathematics Education* 22 (1):3-29.

Cobb, R. , and L.P. Steffe. 1983. The constructivist researcher as teacher and model builder. *Journal for Research in Mathematics Education* 14:83-94.

Cohen, D.K. 1990. A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis* 14:327-345.

Coleman, J.S. , and et al. 1966. Equality of Educational Opportunity. Washington: U.S. Government.

Cuban, L. 1984. *How teachers taught*. New York: Logman.

Davis, R.B. 1984. *Learning mathematics:The cognitive science approach to mathematics education*. Norwood: Ablex.

Fennema, E., and J. Sherman. 1977. Sex-related differences in mathematics achievement, spatial visualization, and affective factors. *American Educational Research Journal* 14:51-71.

Ferrini-Mundy, J., and L. Johnson. 1994. Recognising and recording reform in mathematics: New questions, many answers. *The Mathematics Teacher* 87 (3):190-193.

Fuller, W.A. 1987. *Measurement error models*. New York: John Wiley & Sons.

Hart, J. 1994. Curriculum effort lagging, some say. *Boston Globe*, October 26, 19.

Hiebert, J., ed. 1986. *Conceptual and procedural knowledge: The case of mathematics*. Hillside: Lawrence Erlbaum Associates.

Hiebert, J. , and W. Diana. 1993. Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal* 30:393-425.

Jenks, C. , M. Smith, H. Acland, M.J. Bane, D. Cohen, H. Gintis, B. Heyns, and S. Michelson. 1972. *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic Books.

Kilpatrick, J. 1978. Variables and methodologies in research on problem solving. In *Mathematical problem solving*, edited by L. Hatfield. Columbus: ERIC.

Knapp, M.S., and Associates. 1995. *Teaching for Meaning in High-Poverty Classrooms*. New York: Teachers College Press.

Kupermintz, H., M.E. Michele, L.S. Hamilton, J.T. Talbert, and R.E. Snow. 1995. Enhancing the validity and usefulness of large-scale educational assessments: I. NELS:88 mathematics acheivement. *American Educational Research Journal* 32 (3):555-581.

Laird, N.M. 1992. Estimating rates of change in clinical studies. In *Research Designs and Methods in Psychiatry*, edited by M. Fava and J. F. Rosenbaum: Elsevier Science Publication.

Lampert, M. 1986. Knowing, doing and teaching mathematics. *Cognition and Instruction* 3:305-42.

Lesh, R. , and M. Landau, eds. 1983. *Acquisition of mathematics concepts and processes*. New York: Academic Press.

Marshall, S., S. Fuhrman, and J. O'Day. 1994. National curriculum standards: Are they desirable and feasible? In *The governance of curriculum: 1994 yearbook of the Association for Supervision and Curriculum Development*, edited by R. Elmore and S. Fuhrman. Alexandria: ASCD.

Metz, M.H. 1978. *Classrooms and corridors: The crisis of authority in desegregated secondary schools*. Berkeley: University of California Press.

Miller, J.D., R.W. Suchner, T.B. Hoffer, K.G. Brown, and C. Nelson. 1991. LSAY codebook: Student, parent, and teacher data for longitudinal years one through four (1987-1991). DeKalb: Public Opinion Laboratory, Northern Illinois University.

Mosle, S. 1996. Scoures count: Principal Michael Johnson is helping poor black students pass state exams. So why do school reformers see him as a pariah? *The New York Times Magazine*, September 8, 1996.

National Assessment of Educational Progress. 1986. Math Objectives 1985-1986 Assessment. Princeton: Educational Testing Service.

National Council of Teachers of Mathematics. 1989. Curriculum and evaluation standards for school mathematics. Reston: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. 1991. Professional standards for teaching mathematics. Reston: National Council of Teachers of Mathematics.

National Research Council. 1989. Everybody counts: A reprot on the future of mathematics education. Washington, D.C.: National Research Council.

Oakes, J. 1985. *Keeping track: How schools structure inequality*. New Haven: Yale University Press.

Pelavin, S.H., and M.K. Kane. 1990. Changing the odds: Factors increasing access to college. New York: The College Entrance Examination Board.

Porter, A.C. , M.W. Kirst, E.J. Osthoff, J.L. Smithson, and S.A. Schneider. 1993. Reform up close: An analysis of high school mathematics and science classrooms. Madison: Wisconsin Center for Education Research.

Raudenbush, S.W. , B. Rowan, and Y.F. Cheong. 1993. Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal* 30: 523-553.

Resnick, L.B., and D.P. Resnick. 1991. Assessing the thinking curriculum: Net tools for educational reform. In *Changing assessments: Alternative view of aptitude, achievement, and instruction*, edited by B. R. Gifford and M. C. O'Connor. Boston: Kluwer Academic Publishers.

Rogosa, D., and H. Saner. 1995. Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behaviroral Statistics* 20 (2):149-170.

Rogosa, D.R. , D. Brandt, and M. Zimowski. 1982. A growth curve approach to the measurement of change. *Psychological Bulletin* 90:726-748.

Romberg, T.A. , K. Wilson, and M. Khaketla. 1989. An examination of six standardized mathematics tests for grade eight. Madison: Wisconsin Center for Education Research.

Rowan, B. , S.W. Raudenbush, and S.J. Kang. 1991. Organizational design in high schools: A multilevel analysis. *American Journal of Education* 99:238-266.

Schoenfeld, A.H. 1987a. *Cognitive science and mathematics education*. Edited by A. H. Schoenfeld, *In Cognitive science and mathematics education*. Hillside: Lawrence Erlbaum Associates.

Schoenfeld, A.H. 1987b. Cognitive science and mathematics education: An overview. In *In Cognitive science and mathematics education*, edited by A. H. Schoenfeld. Hillside: Lawrence Erlbaum Associates.

Schoenfeld, A.H. 1987c. What's all the fuss about metacognition. In *Cognitive science and mathematics education*, edited by A. H. Schoenfeld. Hillside: Lawrence Erlbaum Associates.

Silver, E. 1987. Foundations of cognitive theory and research for mathematics problem solving. In *In Cognitive science and mathematics education*, edited by A. H. Schoenfeld. Hillside: Lawrence Erlbaum Associates.

Silver, E. , and A.P. Kenney. 1993. An examination of relationships between the 1990 NAEP mathematics items for grade 8 and selected themes from the NCTM standards. *Journal for Research in Mathematics Education* 24:159-67.

Shulman, L.S. 1986. Paradigms and research programs in the study of teaching: A contemporary perspective. In *Handbook of research on teaching*, edited by M. Wittrock. New York: Macmillan.

Simon, M.A., and D. Schifter. 1993. Toward a constructivist perspective: The impact of a mathematics teacher inservice program on students. *Educational Studies in Mathematics* 25:331-340.

Stata Corporation. 1993. *Stata reference manuel realeas 3.1*. 6 ed. College Station.

Stevenson, D.L., K.S. Schiller, and B. Schneider. 1994. National sequences of opportunities for learning. *Sociology of Education* 67:184-198.

White, B. 1993. Inner-city schools find harder classes really get results. *The Atlanta Journal and Constitution*, December 12, 4.

Wiley, D.E., and B. Yoon. 1995. Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis* 17 (3):355-370.

Endnotes

[1] Values were not imputed because most of the data were missing due to teachers not completing their classroom surveys, and/or students missing test score data.

[2] The variable construction was done at the classroom level in order to prevent the classrooms with more students from holding disproportionate sway. In addition, the lecture and seatwork variables were reverse coded (so that high values indicate infrequent use) to allow for the possibility of creating an index representing the degree of active teaching.

[3] Only one composite in the PCA had an eigenvalue of greater than 1. It contained 49 percent of the variance from the 4 original variables.

[4] This is confirmed by the fact that the Mathematics Objectives Committee charged with the overhaul of the 1990 NAEP utilized the draft version of the *NCTM Curriculum Standards* to develop their objectives for the new exam (Silver & Kenney, 1993).

[5] The impact of these extreme values was explored by setting aside the students and classrooms with very unusual gains, none of the findings presented below were dramatically changed.

[6] It was hypothesized that the relationship between teaching experience and gain scores may not be linear, and a bivariate plot between these variables confirmed that there was some slight curvature in the form of a learning curve. This would make sense if during the initial years teachers experience rapid improvements in gain scores, but with time the degree of improvement slows down. However, it was decided to use it untransformed since neither taking the log of experience, nor fitting it as a quadratic, made its fit on the bivariate plot appear more linear than its untransformed version.

[7] The effect sizes were estimated by multiplying each variable's parameters estimate by its standard deviation and dividing by a standard deviation in gain scores. Teaching style: ES = 1.54(1)/8.77 = .18. Teaching experience: ES = .22(7.93)/8.77 = .18

[8] Style is standardized to have a mean of 0 and a standard deviation of 1. The pre-test has a mean of 63.5 and a standard deviation of 9.2.

BEST COPY AVAILABLE

Table 1: Distributions of responses to questions posed to LSAY teachers about their instructional practices (n=37).

About how much classroom time do you spend on each of the following with this class during a typical week?	Mean	sd	None	30 min	1 hour	2 hours	More than 3 hr.
1)Seatwork	3.24	0.95	2	14	14	5	0
2) Lecturing to the class	2.76	0.72	0	6	16	15	1
3) Leading discussions	2.43	0.80	5	13	17	2	0
4) Students work in small groups	2.14	0.89	9	17	8	3	0

BEST COPY AVAILABLE

Table 2: Descriptive Statistics for Teacher and Student Level Variables

	Mean	Std Dev	Minimum	Maximum
Teacher Variables (n=37)				
Active Teaching	0.00	1.00	-2.32	1.91
Years Teaching	16.46	7.31	1.00	31.00
Masters (0=no, 1=yes)	0.57			
Math Courses	11.74	5.17	3.00	26.00
Student Variables (n=325)				
Gain Score	4.67	8.77	-32.32	48.04
MALE (0=female, 1=male)	0.51			
Parent Push	7.93	1.79	1.33	10.00
SES	0.41	0.68	-1.30	1.81

Table 3: Correlations within the student and teacher level variables.

	Student Variables (n=325)		
	Parent Push	SES	MALE
Gain Score	.02	.07	.09
Parent Push		.09~	.02
SES			.09

	Teacher Variables (n=37)		
	Years Teaching	Math Courses	Masters
Active Teaching	-.05	-.34*	.02
Years Teaching		.01	.25
Math Courses			-.04

Table 4: Taxonomy of models predicting student gains in mathematics. Regression coefficient estimates (with standard errors in parentheses)

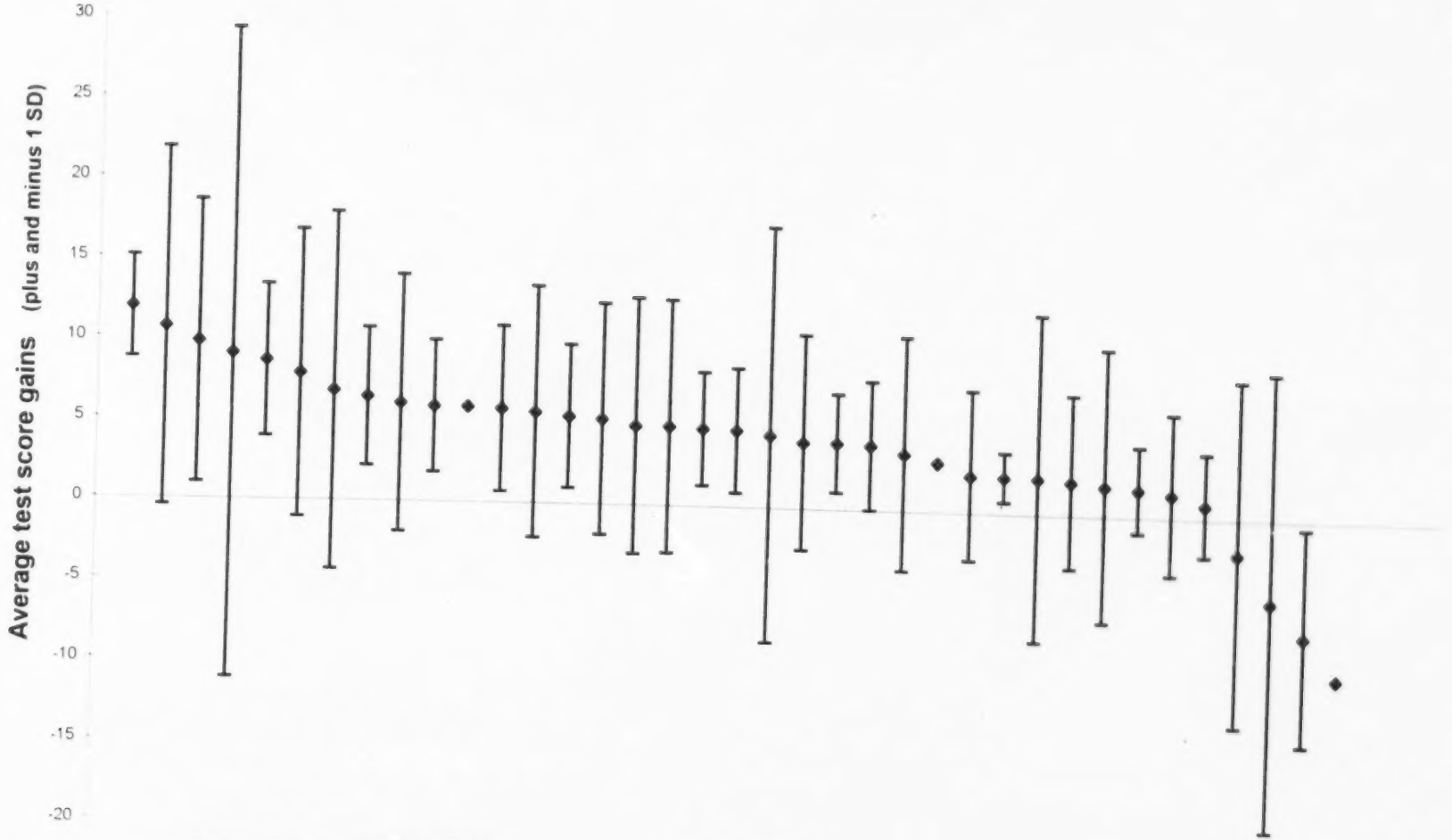
	Models[a]				
Predictors	1	2	3	4	5
Student Level					
Intercept	4.32***	3.35*	4.47***	4.76***	4.76***
	(.67)	(.97)	(.60)	(.48)	(.57)
Parent Push	0.11				
	(.31)				
SES	0.54				
	(.84)				
Male	0.106				
	(1.06)				
Teacher Level					
Years Teaching		0.18~	0.19*	0.22**	0.22**
		(.11)	(.09)	(.07)	(.08)
Math Courses		0.00			
		(.14)			
Masters Degree		0.63			
		(1.52)			
Active Teaching			-1.03~	-1.54*	-1.54~
			(.63)	(.73)	(.91)

~p< .10 *p<.05 **p<.01 ***p<.001

[a]Models 1-3 are two-level HLM models. Model 4 was fit using STATA's errors-in-variables correction. Model 5 presents the HLM "adjusted" standard errors for model 4's reliability corrected parameter estimates.

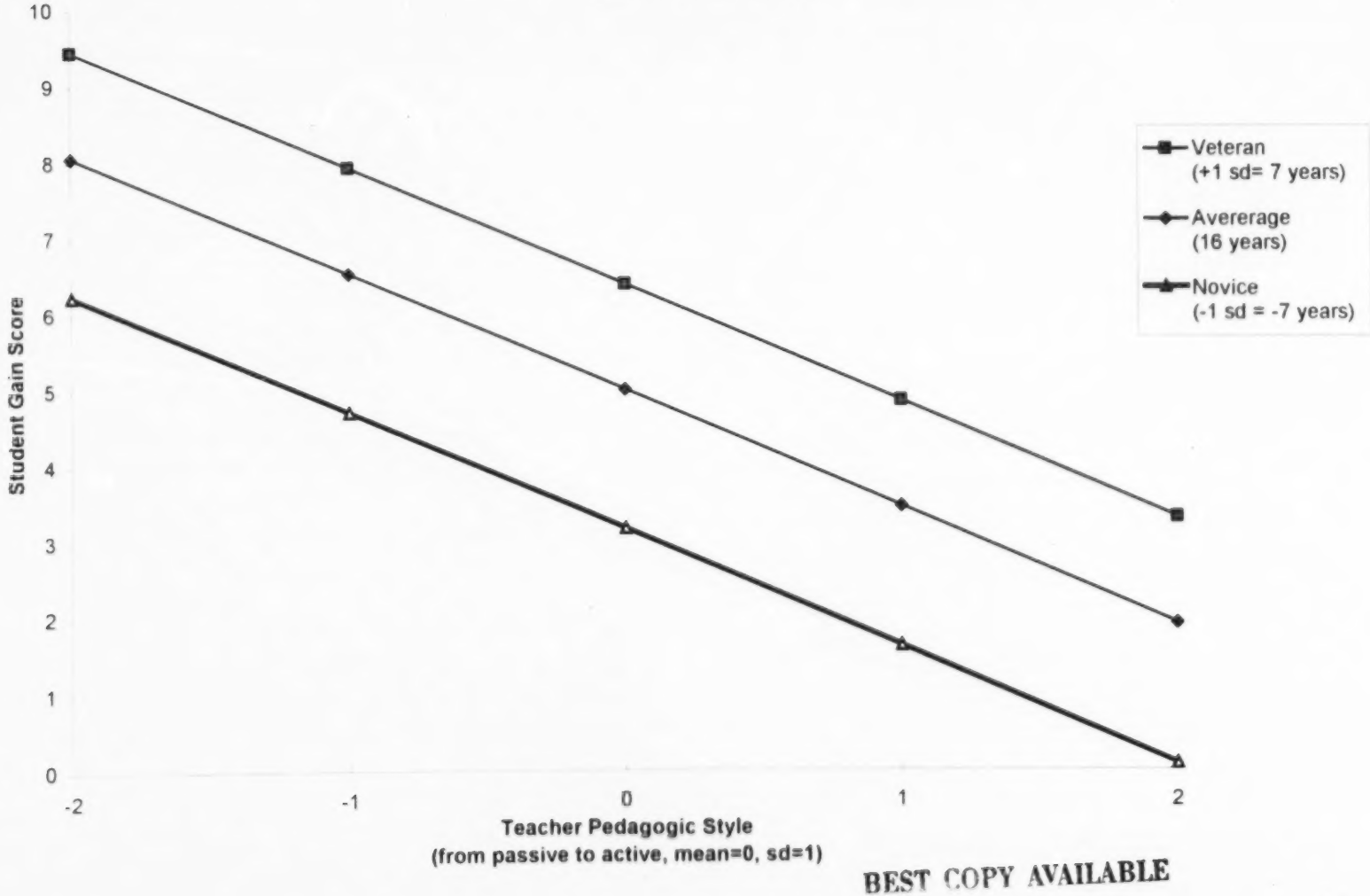
BEST COPY AVAILABLE

Figure 1: Variation in test score gains by classroom



BEST COPY AVAILABLE

Figure 2: Fitted relationship between student gain scores and teacher pedagogic style by level of teaching experience.



BEST COPY AVAILABLE